Challenges of Differentially Private Release of Data Under an Open-world Assumption

Elham Naghizade, James Bailey, Lars Kulik, Egemen Tanin
School of Computing and Information Systems
The University of Melbourne
enaghi,baileyj,lkulik,etanin@unimelb.edu.au

ABSTRACT

Since its introduction a decade ago, differential privacy has been deployed and adapted in different application scenarios due to its rigorous protection of individuals' privacy regardless of the adversary's background knowledge. An urgent open research issue is how to query/release time evolving datasets in a differentially private manner. Most of the proposed solutions in this area focus on releasing private counters or histograms, which involve low sensitivity, and the main focus of these solutions is minimizing the amount of noise and the utility loss throughout the process. In this paper we consider the case of releasing private numerical values with unbounded sensitivity in a dataset that grows over time. While providing utility bounds for such case is of particular interest, we show that straightforward application of current mechanisms cannot guarantee (differential) privacy for individuals under an open-world assumption where data is continuously being updated, especially if the dataset is updated by an outlier.

CCS CONCEPTS

•**Theory of computation** → **Theory and algorithms for application domains;** *Database theory; Theory of database privacy and security;*

KEYWORDS

Dynamic Datasets, Differential Privacy, Unbounded Global Sensitivity, Numerical Queries

1 INTRODUCTION

The growth in information technology and its penetration into our daily life has resulted in an ever-increasing amount of personal data being collected. Harnessing such large and diverse data has created numerous opportunities To provide high quality, personalised services, data exchange has become a key practice. However, a naive exchange of data may lead to significant privacy breaches. This has encouraged a range of studies on how to balance user privacy against data utility. A key goal of data exchange is to derive statistical or aggregate findings about a dataset as this limits privacy implications for a single user.

SSDBM '17, Chicago, IL, USA

DOI: http://dx.doi.org/10.1145/3085504.3085531

K-anonymity [20, 21] and its subsequent variations, such as *l*-diversity [17] and *t*-closeness [16] were among the early proposed solutions to provide *statistical privacy* in data release scenarios. However, under various circumstances such privacy schemes may be breached [7, 9]. This lead to the introduction of Differential Privacy (DP) [4], that aims to preserve users' privacy through guaranteeing that the presence or absence of a user in the dataset does not have a *considerable* effect on the outcome of a query or a data release. DP guarantees what is protected and its amount of protection [13].

Most existing approaches that provide differential privacy focus on a *static* dataset, i.e., they assume a closed world where all data is known and does not change anymore. This makes current DP solutions impractical for real-world scenarios where people constantly produce new data or optout of existing datasets and hence, rapidly changing an existing dataset. The study of differentially private releases of time series or updated outputs has recently gained attention. The authors in [6] propose differentially private counter releases under continual observations, while the work in [2] does not assume any upper bound for the number of releases. The authors in [3] propose a method that can handle dynamic leaves and joins, e.g., nodes in a sensing network. Each node encrypts and perturbs its value and a trusted aggregator would compute the noisy sum of the values. The proposed approach in [15] develops a mechanism to release differentially private histograms in a dynamic setting.

However, the above-mentioned state-of-the-art studies focus on cases with bounded and small global sensitivity, which intuitively models the maximum difference a single user can make wrt a certain query. Consequently, these DP mechanisms do not have to bound sensitivity and rather focus on improving data utility since the noise level accumulates over time. In this paper we focus on the provided privacy levels for the case of differentially private release of numerical values (with unbounded sensitivity), e.g., average salary, average daily power consumption of a household, etc. This is of particular concern when the update is an outlier:

Example 1: A new family has moved to a suburb and takes part in a survey. A similar survey has been conducted before their move and its result has been released in a differentially private manner. Suppose we are interested in a private release of the average salary of families residing in that suburb.

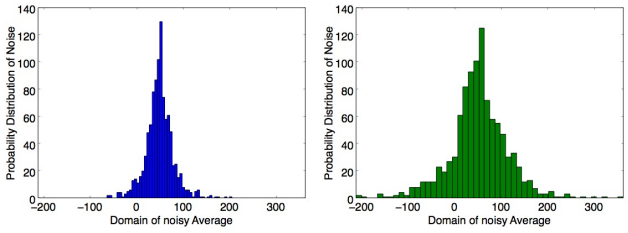
Suppose the dataset in Example 1 is updated with an outlier, e.g., a new family with a considerably higher than average salary moves to the suburb. Figure 1 shows the change in the distribution of differentially private average salaries before and after the update. Computing the amount of change in the noise level, an adversary may infer how much the global sensitivity has changed. We provide more details later in this paper.

The continual release of this information is of great importance in many areas, e.g., urban planning applications. In our work, we show that although the current proposed solutions provide privacy guarantees for a single release, they are vulnerable if an outlier (or a different group of users) is added to the dataset because the prior private release of the dataset or query results can be utilized as background knowledge of an adversary to perform a privacy attack over the next release. To the best of our knowledge our paper is the first study that explores the plausibility of releasing numerical values with differential privacy guarantees under an open world assumption. This calls for a rethink of current DP mechanisms that deal with numerical values of users to ensure user privacy in a dynamic world.

2 RELATED WORK

Recent studies aim to provide differential privacy for scenarios which involve dynamic datasets and multiple releases over time. [6] and [2] proposed methods to release continual DP counters while protecting the presence or absence of an event. [6] introduced the notion of pan privacy, which provides differential privacy guarantees even if the adversary has knowledge of the internal state of the mechanism at a single timestamp. [19] suggests using a combination of encryption and Laplace perturbation to release time series statistics. However, their mechanism needs full knowledge of the stream to perform Discrete Fourier Transformation, which makes it less practical in real world scenarios. Rather than event privacy, the authors in [11] propose a method to provide w-event privacy, i.e., user-privacy over w consecutive timestamps, and allows updates in the values and the size of the dataset does not change. The proposed mechanism in [3] computes a differentially private sum in a distributed system with users joining/leaving the dataset. Likewise, the authors in [15] propose a DP-aware histogram release for dynamic datasets, in which people leave and join the dataset, however, they assume cases where the domain of D remains the same, i.e., bounded sensitivity.

Differential privacy claims that the presence or absence of an individual datum is protected regardless of the adversary's background knowledge. Certain background knowledge of the adversary has recently been proven to breach privacy: access to prior deterministic aggregate information about contingency tables [12], knowledge about the correlation between data records in the datasets [23]. However, having access to the prior differentially private data releases has not been studied in the literature.



(a) Initial noisy salary release. **(b) Noisy salary with an outlier.**

Figure 1: The distribution of the result of an average query when $k = 100$ and $\varepsilon = 0.1$. The average does not change due to adding an outlier, however, the noise level and standard deviation of the distributions change considerably.

3 PRELIMINARIES

3.1 Differential Privacy

Differential privacy was proposed to preserve a user's *evidence of participation* in a dataset. Following this scheme the adversary's knowledge must not *considerably* change with or without the participation of a single user:

Definition 1 [4]*:* A mechanism, $\mathscr{M}$ is ε*-differentially private* over a query function f if for any two neighboring datasets D and D', and any $r \in S$ where $S \subset Range(f)$:

$$\frac{Pr[\mathscr{M}(D) = r]}{Pr[\mathscr{M}(D') = r]} \leq e^{\varepsilon} \ \varepsilon > 0$$

where ε is the privacy budget of the process and along with the *global sensitivity* determines the amount of noise to be added.

Definition 2 [4]*: Global sensitivity* is the maximum distance of the result of the query, $f(.)$, over all possible queries and datasets:

$$\Delta f = \max_{D,D',Q} |f(D) - f(D')|$$

A common way of guaranteeing DP is to add Laplacian noise $\mathscr{L}(\lambda)$ to the true answer of the query where $\lambda = \Delta f/\varepsilon$.

3.2 Unbounded sensitivity of Numerical Releases

Based on its definition, Δf must be determined according to all the observations in the original dataset as well as all potential samples that are not in the dataset (as it needs to protect users who did not participate in the dataset as well). This may lead to an unbounded global sensitivity, which has not been addressed in the literature. For instance, if a dataset, D_s consists of the salary of residents in a given suburb, one may argue that the records have a lower bound of zero (someone with no salary), however, the upper bound needs to be selected in order to protect even people who are not residing in that suburb and hence are not in that dataset. A very delicate question is what should be chosen as an upper bound for Δf since simply adding $\mathscr{L}(\infty)$ may result in no utility due to the considerable increase in the noise to signal ratio.

3.2.1 Maximizing Privacy. Following Example 1 assume the actual average salary in D is $\hat{x}_a$. When determining an upper bound for Δf, we have to consider an individual with salary s who is not in the sample dataset, but needs to be protected. Adding such a virtual individual to D with size n, incurs an approximate relative error of $e = \frac{w - \hat{x}_a}{n\hat{x}_a}$ to the original average value, i.e., $\hat{x}'_a \approx \hat{x}_a(1+e)$.

This means that within a suburb where $n = 15000$ and $\hat{x}_a = \$35,000$, assuming a (virtual) salary, s, of around 5 million dollars as s moves the average by 1%, and a salary of 0.5 billion dollars (still a possible salary) doubles the average, rendering no data utility. Note that up to this stage no noise has been added to the result and the only source of error is due to determining Δf in a data-independent manner.

Traditionally, the numerical values are bounded by a potential maximum and global sensitivity is computed according to that [5]. We choose to bound the estimated range of salaries as $S = [0, m\hat{x}_a]$ ($m \in \mathbb{N}, m > 1$) as proposed in [10]. When the size of the dataset is publicly known, this implies adding noise proportional to $\mathscr{L}(\frac{m\hat{x}_a)}{n\varepsilon})$ to the sum of the salaries. In such case the utility loss is linearly dependent to the estimated range of salaries and the privacy budget.

While the data owner my prefer larger ranges (larger m) to preserve privacy, the utility of the query results deteriorates considerably.

3.2.2 Maintaining Utility. In this paper, we assume the release of query results is rewarded by data analyst(s) under the condition that it meets certain utility requirements.

Definition 3 (a,δ)-loss bound: The *query result, r is bounded by* (a,δ) if the probability of r deviating a% from the true value is less than δ:

$$Pr(|r-\hat{x}| \geqslant a\hat{x}) \leq \delta$$

We can rewrite the above expression using the Chebyshev inequality. Having $\sigma = \sqrt{2}\frac{m\hat{x}}{n\varepsilon}$ and a as the maximum relative error tolerated:

$$Pr(|r-\hat{x}| \geqslant a\hat{x}) = Pr(|r-\hat{x}| \geqslant k\sigma) \leq \frac{1}{k^2}$$

In a case where the data owner wants to determine m arbitrarily, this gives a lower bound on choosing ε. Having $\frac{1}{k^2} \leq \delta \rightarrow k \geqslant \frac{1}{\sqrt{\delta}}$ and setting $k\sigma \leq a\hat{x}$, we reach $\varepsilon \geqslant \sqrt{\frac{2}{\delta}}\frac{m}{na}$. A similar computation can determine m for certain εs.

4 PROBLEM DESCRIPTION

We consider the case where the data owner releases multiple differentially private query answers/datasets. This is particularly necessary in time-evolving datasets where users are dynamically leaving or joining the dataset and studying the historical behavior of the users is of importance.

4.1 Incremental continual release of numerical values

In our scenario, the data curator (potential adversary) may access the dataset N times. This may be in the form of asking queries, i.e., an interactive model, or gaining access to a private release of the dataset, i.e., a non-interactive model.

Definition 4: An *incrementally evolving dataset*, $\mathscr{D}$ is a set of size N where $\forall D_i, D_{i+1} \in \mathscr{D}$ we have $|D_{i+1}| = |D_i| + 1$ and $D_i \subset D_{i+1}$, $\forall i \in \{1,\ldots,N-1\}$.

Upon receiving an update (it can be modified to cover k updates), the data owner wants to release a private average of values for D_{i+1} (or D_{i+k}). A straightforward approach towards releasing differentially private results for each update is to uniformly divide the overall privacy budget, $\mathscr{E}$, and add independent noise with $\varepsilon = \mathscr{E}/N$ to the true values each time a query is received. Such an approach is supposed to ultimately guarantee εN-differential privacy according to the sequence composition theorem [18] since each sequence of the computation has ε-differential privacy.

To simplify our discussion in this section we assume that the values are monotonically increasing. Note that as the system is dynamic, D_i and D_{i+1}, $\forall 1 \leq i \leq t$ may not necessarily have the same number of users. Intuitively, the amount of temporal gaps between each access reflects to what extent these datasets are different. Without loss of generality, we assume each update is followed by a query from the curator and hence a noisy release is observed.

4.2 Privacy risks of continual releases

As established in the previous section, bounding global sensitivity in a data independent way would adversely affect the output utility. We show now that determining the global sensitivity in a data dependent manner may lead to a breach of privacy.

Suppose r_i and r_{i+1} are the differentially private results to the average query at two consecutive timestamps, t_i and t_{i+1} where D_i and D_{i+1} are two incremental neighbor datasets. Following the mechanism in [10], $r = \hat{x} + \mathscr{L}(\frac{m\hat{x}}{n\varepsilon})$ and thus $Pr(\mathscr{M}(D) = r) = \frac{n\varepsilon}{2m\hat{x}} e^{\frac{-n\varepsilon}{m\hat{x}}|r-\hat{x}|}$ where $r \in \mathbb{R}^+$ in our scenario.

We show through the following example that such approach does not necessarily guarantee differential privacy for a new user joining the dataset.

We have $c = \frac{\hat{x}_{i+1}}{\hat{x}_i}$ where $c > 1$ (i.e., we assume that the average is increasing to simplify the discussion). We also assume the datasets are large enough, hence $n_i \approx n_{i+1}$. As $|D_{i+1} - D_i| = 1$, based on Definition 1 we enjoy $(i+1)\varepsilon$-differential privacy if:

$$\frac{Pr[\mathscr{M}(D_i) = r_i]}{Pr[\mathscr{M}(D_{i+1}) = r_{i+1}]} \leq e^{\varepsilon(i+1)}$$

Without loss of generality we can assume $N = 2$ (the general case for k follows by induction). Since $\Delta f_1 = m\hat{x_1}/n_1$ and $\Delta f_2 = m(c\hat{x_1})/n_1$, for r_1 and r_2 following a Laplacian distribution, we obtain:

$$e^{-2\varepsilon} \leq \frac{\varepsilon n_1}{2\Delta f_1} e^{\frac{-\varepsilon n_1 |r_1 - \hat{x_1}|}{\Delta f_1}} \Big/ \frac{\varepsilon n_1}{2c\Delta f_1} e^{\frac{-\varepsilon n_1 |r_2 - c\hat{x_1}|}{c\Delta f_1}} \leq e^{2\varepsilon}$$

Rewriting e^a/e^b as $e^{(a-b)}$ and applying a logarithm to the inequality we obtain:

$$-2\varepsilon \leq ln(c) - \frac{\varepsilon n_1}{m\hat{x_1}}(|r_1 - \hat{x_1}| - 1/c|r_2 - c\hat{x_1}|) \leq 2\varepsilon$$

Based on its definition, differential privacy must protect the evidence of presence of the newly added user in D_2, i.e., $\forall (r_1, r_2) \in S \times S$. Now suppose $S = [\sigma_2, c\hat{x_1} + c\sigma_2]$. This means for the (highly unlikely yet possible) worst case scenario when $r_1 = \sigma_2, r_2 = c\hat{x_1} + c\sigma_2$, we require to have a privacy budget of $2\varepsilon \geqslant ln(c)$, however, for any arbitrarily chosen $\varepsilon \leq ln(c)/2$ the 2ε-differential privacy will not be satisfied.

The above example shows that ε-differential privacy ε cannot be selected arbitrarily. Moreover, it is not possible to predetermine it (as in the case of uniformly assigning ε_t to each release) since it depends to the amount of change incurred as a result of adding a new user. This is of particular concern if the new record is an outlier with respect to the previous records in the dataset. For instance, the new family in the first scenario have a considerably higher salary than the normal salary range in that suburb (a usual case of gentrification). In such a case, having the initial differentially private query answers as the background knowledge, we not only may provide *evidence of participation* for the new record, but we may also estimate the value of the record.

5 PROPOSED INFERENCE ATTACK ON CONTINUAL RELEASES

Generally, it is not possible to simply observe a changed average value for a successful attack to a differentially but evolving private dataset. The effect of adding an outlier to the dataset is smoothed out by a growing dataset. Further, adding noise proportional to the size of the population ensures small deviations from the true average results for large datasets. Consequently, an adversary is normally not able to detect an outlier simply from the fluctuations in the average results. When an outlier is added to the dataset, the change

in the noisy released average is negligible. However, having full knowledge about the DP mechanism, the adversary knows that the noise level is data dependent at each release and is able to use *the sequence of differentially private average results as background knowledge for an attack*.

Having the sequence of query results over time $\mathscr{R} = \{r_1, r_2, \ldots, r_N\}$, the mean and standard deviation of $\mathscr{R}$ may be estimated using the maximum likelihood estimators (MLE). Since $r = \hat{x} + Lap(\lambda)$, the estimated mean, $\hat{\mu}$ and $\hat{\sigma}$ may be inferred as the true average of the population and the noise level, respectively. For a sample of random variables following a Laplace distribution, $\hat{\mu}$ is estimated as the median of the sample and $\hat{\sigma}$ is estimated as the mean of the distance of the samples from $\hat{\mu}$, i.e., $\hat{\sigma} = \frac{1}{N}\sum_{i=1}^{N}|r_i - \hat{\mu}|$. Moreover, for any sample we can estimate $[2n\hat{\sigma}/\tilde{\chi}^2_{1-\frac{\alpha}{2}}(v) < \sigma < 2n\hat{\sigma}/\tilde{\chi}^2_{\frac{\alpha}{2}}(v)]$ with $1-\alpha$ confidence where $\tilde{\chi}^2(v)$ is a chi square distribution with v degree of freedom and v is proportional to the size of the sample and is equal to $2E(\frac{\hat{\sigma}}{\sigma})$ [1]. Note that for any $R_{wi} \in \mathscr{R}_w$, if there is no overlap between the confidence intervals of R_{wi} and its immediate neighbors the adversary infers with 100% confidence that a person with an anomalous value has participated in the i_{th} window. However, our experiments show that such cases are rare as the outlier needs to be significantly larger than the rest of the population as the added noise blurs the borders.

Data: $\mathscr{R}, w, \delta_s$.
Result: ID of anomalous segments.
$ID_{out} \leftarrow []$;
$R_w \leftarrow Segment(\mathscr{R}, w)$;
while $1 < i < N/w$ **do**
  $\hat{\theta}_i \leftarrow MLE(R_i^w)$;
  $Score_{i+1} \leftarrow p(R_{i+1}|\hat{\theta}_i)$;
  if $Score_{i+1} > \delta_s$ **then**
    $ID_{out} \leftarrow i+1$;
  end
  $i \leftarrow i+1$;
end

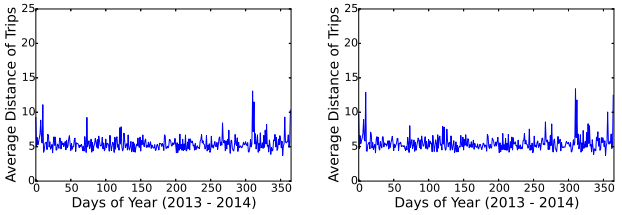
Algorithm 1: Inference Attack on Naive Releases

Algorithm 1 shows a potential inference attack that can be performed by an adversary to detect an outlier regardless of the noise added to blur the difference. After segmenting the released noisy results, the process mainly compromise of two steps: i) fitting an estimator to the i_{th} set of observations, i.e., $\hat{\theta}_i$ and ii) checking how likely is to observe the next segment based on $\hat{\theta}_i$. In our work we used MLE as the adversary have full knowledge of the noise function, otherwise it is possible to utilize other estimators, such as the models used in [8] to find $\hat{\theta}_i$. Moreover, similar to [22], we use the log likelihood probability as our scoring function:

$$Score(R_{i+1}) = -log\prod_{i=1}^{w} L(R_{i+1}|\hat{\theta}_i)$$

The higher the score, the less likely it is that R_{i+1} is drawn from the same distribution.

Definition 3: A $w-$segmented sequence $\mathscr{R}_w$ is a discrete set of segments of noisy results with length w derived from $\mathscr{R}$, i.e., $\mathscr{R}_w = \{R_{w1}, R_{w2}, ..., R_{wk}\}$ where $R_{wi} = (r_{(i-1)w+1}, r_{(i-1)w+2}, .., r_{iw})$ for $1 \leq i \leq k$.



(a) $\hat{d}$ in real dataset. **(b) $\hat{d}$ in synthetic dataset.**

Figure 2: The average distance ($\hat{d}$) of daily trips in Porto taxi trajectory in real dataset and synthetic dataset with $n = 1000$.

Two different strategies are adopted to estimate $\hat{\theta}_i$: $\hat{\theta}_i$ can be estimated using either the i_{th} set of observations, i.e., a fixed number w of observations from the i_{th} segment, denoted as *Fixed Window Estimation (FWE)*. Alternatively, to estimate $\hat{\theta}_i$ the set of observations can be expanded at each step to include all the previous segments, i.e., $\{R_1, ..., R_i\}$, denoted as *Growing Window Estimation (GWE)*.

In addition to *FWE* and *GWE* we employ Levene's test [14] as a statistical inference method that aims to test whether or not the variance of multiple observations are equal (null hypothesis is the equality of variances). Unlike other statistical tests, Levene's test does not assume normality for the underlying distribution of the observations, which makes it an appropriate method to be adopted in our case.

Therefore, if the resulting p-value of Levene's test is less than some significance level, it is unlikely that the observations are sampled from a population with an equal variance. For any two consecutive point in the set of segmented observations, we score the difference between them as $1-p$-value. The less significant the equality of variances, the more likely that the noise level (the standard deviation of Laplace distribution) has been adjusted due to the existence of an outlier.

Such inference attack is of great importance in the privacy community as it may seem a safe practice to add independent noise to each release. In the case of Example 1, the adversary may only know the the new family is much richer than the average residents in the area, however after releasing the second private average an attacker would (i) be able to confidently say if the family has participated in the survey or not and (ii) be able to estimate their real salary within a certain confidence. As a result, a naive release approach cannot guarantee εN-differential privacy over N releases.

6 EXPERIMENTS

To evaluate the performance of our inference attack, we have used a collection of real and synthetic datasets. We used the Porto taxi trajectory dataset. We sampled 19,925 random trajectories within a $10km \times 10km$ area from the Porto taxi trajectory dataset. We consider the scenario of releasing differentially private average distance of trips per day, i.e., N=365. On average there are 55 daily trips in the dataset and the average distance of all trips is $5.45km$. We considered the top 0.001 percentile of the distances ($d > 71.41$) as outliers, which resulted in 20 outlying trips occurring in 19 days.

	Functionality	Range	Default
n	Number of Daily Records	$\approx 55 - 10^4$	10^3
w	Segment Size	$5 - 20$	10
ε	Privacy Parameter	$0.05 - 2.5$	1

Table 1: Experimental settings.

We also created a synthetic dataset with varying number of daily trips: we fit a Gaussian model to the original daily trip distances of the Porto trajectory data and generate two synthetic datasets with 100, 1000, and 10,000 daily trips. Similarly, the top 0.001% of the synthetic distances are classified as outliers, which results in 20, 24 and 24 outlying segments respectively. The synthetic datasets provide the opportunity to explore the performance of our attack in difficult scenarios, where the outlier's effect in the average value becomes insignificant. Figure 2 shows the average number of trips and the real and synthetic average distances over a year.

We examine the performance of our inference attack using the precision (P) and recall (R), where precision shows how many of the detected outlying segments have actually an outlier value in their daily trips and recall specifies how many of the actual outlying segments have been detected by the model. The results are averaged over 100 runs.

As can be seen in Figure 3a, small privacy budgets, i.e., $\varepsilon = 0.5$, incur large amounts of noise in the query results. The small number of daily trips in the Porto taxi dataset significantly increases the noise level (as discussed earlier noise level is proportional to $\frac{\Delta f}{n_i \varepsilon}$ and the smaller the n_i the larger the noise scale). Hence, we consider $\varepsilon > 1$, however when $m = 2$, this setting approximately guarantees a (0.15,0.15)-loss bound (a relatively modest utility bound).

In order to provide comprehensive results, we have considered different settings for our experiments. Table 1 shows a summary of our experimental setting where we evaluated our inference strategies to detect outlying segments in the Porto with respect to the size of the dataset (daily trips), the privacy budget and the size of the segments.

6.1 The Effect of Privacy Budget

We ran the inference attacks for varying εs between 0.05 to 2.5 and Table 2 shows the effect of privacy budget on our inference success. For any ε, *GWE* has the best performance among our scoring strategies since it is using the entire set of previous releases to score a new segments, i.e., it maximizes the use of available background knowledge.

It is expected that larger amounts of ε would result in a more successful inference attack since the amount of noise decreases. This assumption is validated for *GWE* where the increase in ε results in almost 10% of improvement in both precision and recall. However, the precision and recall does not significantly improve when increasing the budget from 1 to 2.5. This is of importance since the data owner is able to determine an upper bound to the success of an adversary in detecting outliers and hence can flexibly opt for larger privacy budgets that considerably improve utility.

Levene's test has its best performance for moderate noise levels. This may be due to the fact that although a small privacy budget adds a large amount of noise to the true value, a large budget imposes a negligible amount of noise, which makes it difficult to determine whether the variance of the segments has changed significantly.

ε	0.05		0.5		1.0		2.5	
Performance	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Levene's Test	0.57	0.52	0.61	0.55	0.61	0.55	0.60	0.53
FWE	0.63	0.56	0.65	0.58	0.65	0.58	0.63	0.57
GWE	0.72	0.64	0.80	0.71	0.80	0.72	0.82	0.73

Table 2: The effect of varying ε on the performance of different outlier detection strategies.

w	5		10		15		20	
Performance	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Levene's Test	0.35	0.57	0.61	0.55	0.75	0.50	0.78	0.47
FWE	0.40	0.63	0.65	0.58	0.79	0.52	1.0	0.60
GWE	0.46	0.65	0.80	0.71	0.91	0.60	1.0	0.60

Table 3: The effect of varying w on the performance of different outlier detection strategies.

6.2 The Effect of Window Size

The increase in the size of the segments would imply larger number of observations to estimate the mean and standard deviations of a sample. The results shown in Table 3 demonstrate that as the size of the segments increase, the precision of all three strategies improves, especially for *GWE* where we witness a sharp rise in its precision when w becomes 10. For $w = 20$, both *FWE* and *GWE* have the maximum precision, i.e., all of the detected outlying segments have actually an outlier in them.

However, recall does not have the same trend as with the increase in the size of the segments, there is a slight decrease in recall for all of the three strategies. This may be due the fact that larger segment sizes decrease the number of total segments in general, and thus the proportion of outlying segments increases. Therefore, missing an outlying segment is penalized harder.

6.3 The Effect of Size of the Dataset

We evaluate the performance of *FWE*, *GWE* and Levene's test against the increase in the size of the dataset (Table 4). Since the increase in n blurs the effect of an outlier in the true average value, we expect the performance of our inference attacks to deteriorate for larger ns. However, our experiments show that an initial increase in the number of daily trips improves the performance of our inference strategies. Although it may sound counterintuitive, this is happening due to i) a decrease in the noise levels and ii) detectability of outliers in datasets with relatively small sizes, e.g., $n = 1000$: Larger ns reduce the noise level considerably, which makes the noisy average releases closer to the actual average value. On the other hand, the actual average values, and respectively the noisy averages, cannot blur the effect of an outlier when n is not large enough.

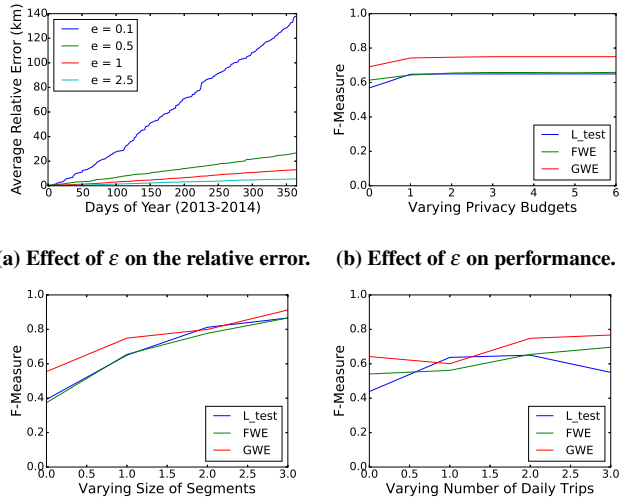
A dataset of 10000 daily trips, worsens the performance of our inference strategies, where Levene's test is most affected by the increase in the dataset size.

7 FINDINGS AND DISCUSSION

We introduced three strategies to detect the outlying segments: *FWE* scores each segment on noisy values based on the most recently observed segment, whereas *GWE* considers all of the previous observations when evaluating the current observation. We also adopt the *Levene's test* that is a statistical inference approach to examines how

n per day	≈ 55		100		1000		10000	
Performance	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Levene's Test	0.44	0.50	0.61	0.65	0.61	0.55	0.52	0.43
FWE	0.55	0.61	0.53	0.56	0.65	0.58	0.67	0.57
GWE	0.62	0.69	0.60	0.64	0.80	0.72	0.77	0.67

Table 4: The effect of varying n on the performance of different outlier detection strategies.



(a) Effect of ε on the relative error. **(b) Effect of ε on performance.**

(c) Effect of w on performance. **(d) Effect of n on performance.**

Figure 3: Effect of varying parameters on performance (determined based on F-Measure) of the inference attacks.

equal the variances of a set of observations are. Our finding suggests that the *GWE* has the best performance with regard to detecting the outlying segment, which is due to the fact that it maximizes the use of available background knowledge (in contrast to *FWE*) to estimate the probability distribution of noisy releases. On the other hand, *Levene's test* is largely unsuccessful to detect the outlying segments, although it follows the intuition behind the attack, i.e., change in the noise scale, to detect the outlying segment. A noteworthy observation of the experimental results is the fact that an increase in the privacy budget cannot improve the performance of our inference attacks from a certain threshold. This knowledge can be used to model the probability of success in an attack to choose a utility-aware budget, hence balancing privacy and utility.

Finally, our inference attacks can be further explored with respect to properties such as the number of the outliers in the dataset and the size of outliers, i.e., how large it is compared to the average values. Moreover, the pace of changes in the dataset is expected to have an effect on the success of our inference attack. The Porto taxi dataset is a highly dynamic dataset, wherein *overlapping* daily trips may not be a widespread phenomenon. However, in a less dynamic dataset where the majority of the underlying users remains untouched, the effect of an outlier may be more readily detectable.

8 CONCLUSION

We focused on the problem of incremental release of differentially private numerical values in a time evolving dataset. Assuming a *safe* yet *utility-aware* upper bound for the maximum possible value and tailoring noise level based on that is a common practice. However, we provided formal and experimental evidence that such mechanism cannot guarantee (differential) privacy over time when data utility is paramount, particularly if the dataset is updated by an outlier. Our finding urges a rethink of straightforward DP mechanisms prior to applying it to complex, time evolving datasets, which are common in many emerging areas such as participatory sensing.

9 ACKNOWLEDGMENT

This research was partially supported under Australian Research Council's Discovery Projects funding scheme (DP170102472).

REFERENCES

[1] L. J. Bain and M. Engelhardt. Interval estimation for the two-parameter double exponential distribution. *Technometrics*, 15(4):875–887, 1973.
[2] T.-H. H. Chan, E. Shi, and D. Song. Private and continual release of statistics. *TISSEC*, 14(3):26:1–26:24, 2011.
[3] T.-H. H. Chan, E. Shi, and D. Song. Privacy-preserving stream aggregation with fault tolerance. In *16th International Conference on Financial Cryptography and Data Security*, 2012.
[4] C. Dwork. *33rd International Colloquium on Automata, Languages and Programming*, chapter Differential Privacy. Springer Berlin Heidelberg, 2006.
[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Third Theory of Cryptography Conference*, 2006.
[6] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. In *ACM Symposium on Theory of Computing*, 2010.
[7] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In *14th ACM International Conference on Knowledge Discovery and Data Mining*, 2008.
[8] V. Guralnik and J. Srivastava. Event detection from time series data. In *5th ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 1999.
[9] M. Hay, G. Miklau, D. Jensen, D. Towsley, and C. Li. Resisting structural re-identification in anonymized social networks. *The VLDB Journal*, 19(6):797–823, 2010.
[10] O. Heffetz and K. Ligett. Privacy and data-based research. Working Paper 19433, National Bureau of Economic Research, September 2013.
[11] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias. Differentially private event sequences over infinite streams. *Proc. VLDB Endow.*, 7(12), 2014.
[12] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2011. ACM.
[13] D. Kifer and A. v. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1), 2014.
[14] M.-P. Lai. *Robust tests on the equality of variances*. PhD thesis, 1997.
[15] H. Li, L. Xiong, X. Jiang, and J. Liu. Differentially private histogram publication for dynamic datasets: an adaptive sampling approach. In *24th ACM International Conference on Information and Knowledge Management*, 2015.
[16] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering*. IEEE, 2007.
[17] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.
[18] F. D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *ACM SIGMOD International Conference on Management of Data*, pages 19–30. ACM, 2009.
[19] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *2010 ACM SIGMOD*. ACM, 2010.
[20] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Technical report, SRI International, 1998.
[21] L. Sweeney. k-anonymity: A Model For Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):1–14, 2002.
[22] K. Yamanishi and J.-i. Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *8th ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 2002.
[23] B. Yang, I. Sato, and H. Nakagawa. Bayesian differential privacy on correlated data. In *ACM SIGMOD International Conference on Management of Data*. ACM, 2015.